

Statistical Alignment and Machine Translation

Berlin Chen

Department of Computer Science & Information Engineering
National Taiwan Normal University

References:

1. Foundations of Statistical Natural Language Processing, Chapter 13
2. Speech and Language Processing (3rd ed.), Chapter 13

Language Divergences and Typology

- There are about 7,000 languages in the world
 - Structural linguistic universals:
 - Every language seems to have nouns and verbs, has ways to ask questions, or issue commands, has linguistic mechanisms for indicating agreement or disagreement
 - Translation divergence: however, when building machine translation (MT) systems We often distinguish the idiosyncratic (獨特的) and lexical differences that must be dealt with one by one

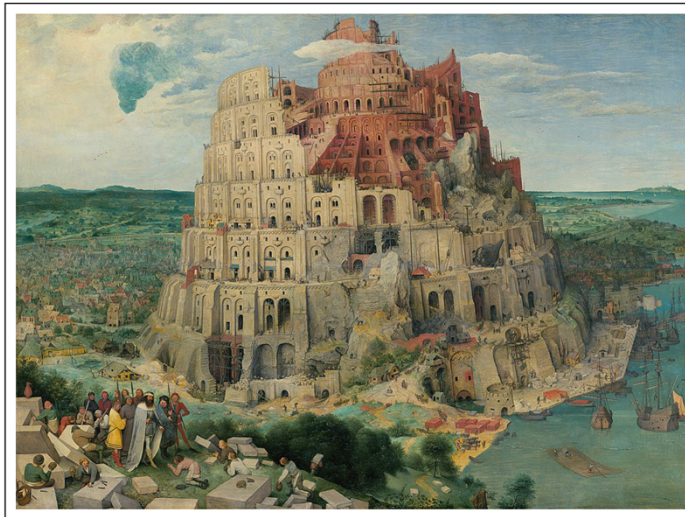


Figure 13.1 The Tower of Babel, Pieter Bruegel 1563. Wikimedia Commons, from the Kunsthistorisches Museum, Vienna.

Word Order Typology (1/2)

- For simple declarative clauses
 - German, French, English, and Mandarin, for example, are all SVO (Subject-Verb-Object) languages
 - Hindi and Japanese, by contrast, are SOV languages, meaning that the verb tends to come at the end of basic clauses
 - Irish and Arabic are VSO languages

English: *He wrote a letter to a friend*

Japanese: *tomodachi ni tegami-o kaita*
friend to letter wrote

Arabic: *katabt risāla li šadq*
wrote letter to friend

SVO languages generally have prepositions, whereas SOV languages generally have postpositions.

Word Order Typology (2/2)

- Word order differences between languages can cause problems for translation, requiring the system to do huge **structural re-orderings** as it generates the output

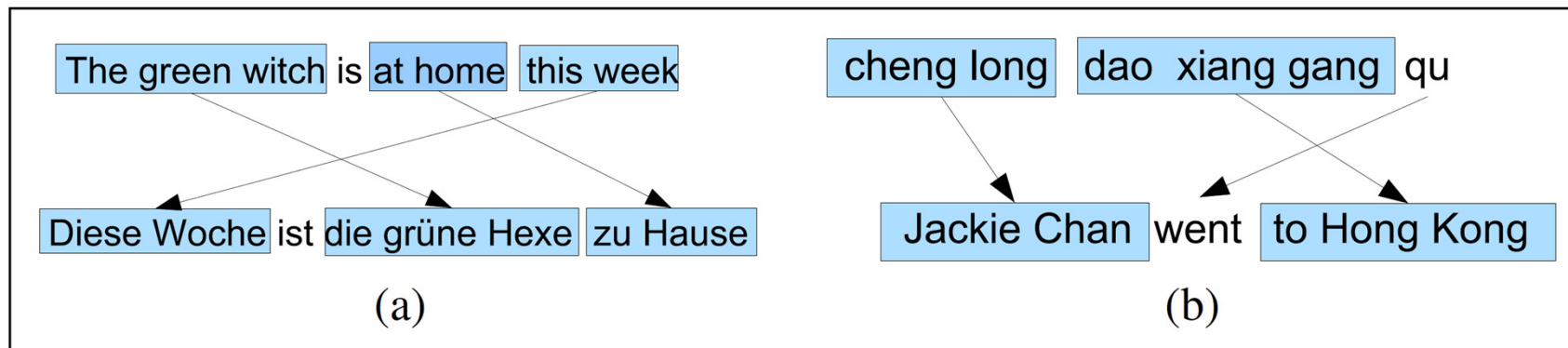


Figure 13.2 Examples of other word order differences: (a) In German, adverbs occur in initial position that in English are more natural later, and tensed verbs occur in second position. (b) In Mandarin, preposition phrases expressing goals often occur pre-verbally, unlike in English.

Lexical Divergences (1/2)

- For any translation, the appropriate word can vary depending on the context, for example:
 - The English source-language word *bass*, for example, can appear in Spanish as the fish *lubina* (歐洲海鱸魚) or the musical instrument *bajo* (低音樂器)
 - For the The English source-language word *wall*, German uses two distinct words: *Wand* for walls inside a building, and *Mauer* for walls outside a building
- Perform translation from would require a kind of specialization, disambiguating the different uses of a word
- The fields of MT and word sense disambiguation (WSD) are closely linked

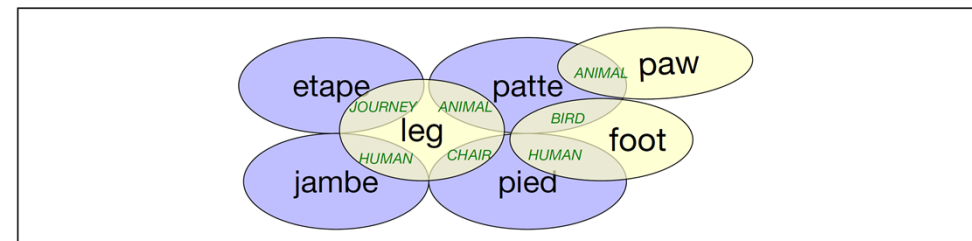


Figure 13.3 The complex overlap between English *leg*, *foot*, etc., and various French translations as discussed by Hutchins and Somers (1992).

Lexical Divergences (2/2)

(Talmy 1991, Slobin 1996)

- **Verb-framed languages:** mark the direction of motion on the verb (leaving the satellites to mark the manner of motion)
 - Languages like Japanese, Tamil, and the many languages in the Romance, Semitic, and Mayan languages families, are **verb-framed**
 - Chinese as well as non-Romance Indo-European languages like English, Swedish, Russian, Hindi, and Farsi are **satellite-framed**

English: *The bottle floated out.*

Spanish: La botella salió flotando.

The bottle exited floating.

A bottle floating out of a cave would be described in English with the direction marked on the particle out, while in Spanish the direction would be marked on the verb.

Morphological Typology

- Morphologically, languages are often characterized along two dimensions of variation
 - The number of morphemes per word
 - **Isolating languages** like Vietnamese and Cantonese, in which each word generally has one morpheme
 - **Polysynthetic languages** like Siberian Yupik (“Eskimo”), in which a single word may have very many morphemes, corresponding to a whole sentence in English
 - The degree to which morphemes are segmentable
 - **Agglutinative (黏結的) languages** like Turkish, in which morphemes have relatively clean boundaries
 - fusion languages like Russian, in which a single affix may conflate (合併) multiple morphemes

Referential Density

- Some languages, like English, require that we use an explicit pronoun when talking about a referent that is given in the discourse
- In other languages like Chinese , Japanese and Spanish, we, however, can sometimes omit pronouns altogether
 - Languages that can omit pronouns are called **pro-drop languages**
- We say that languages that tend to use more pronouns are more **referentially density** dense than those that use more zeros
 - Referentially sparse languages, like Chinese or Japanese, that require the hearer to do more inferential work to recover antecedents are also called **cold languages**
 - Languages that are more explicit and make it easier for the hearer are called **hot languages**

Machine Translation (MT)

- Definition
 - Automatic translation of text or speech from one language to another
- Goal
 - Produce close to error-free output that reads fluently in the target language
 - *Far from it? Or, a solved problem?*
- Current Status
 - Existing systems perform well in restricted domains
 - E.g. weather reports
 - A mix of probabilistic and non-probabilistic components

Issues

- Build high-quality semantic-based MT systems in circumscribed domains
- Abandon automatic MT, build software to assist human translators instead
 - Post-edit the output of a buggy translation
- Develop automatic knowledge acquisition techniques for improving general-purpose MT
 - Supervised or unsupervised learning

Different Strategies for MT

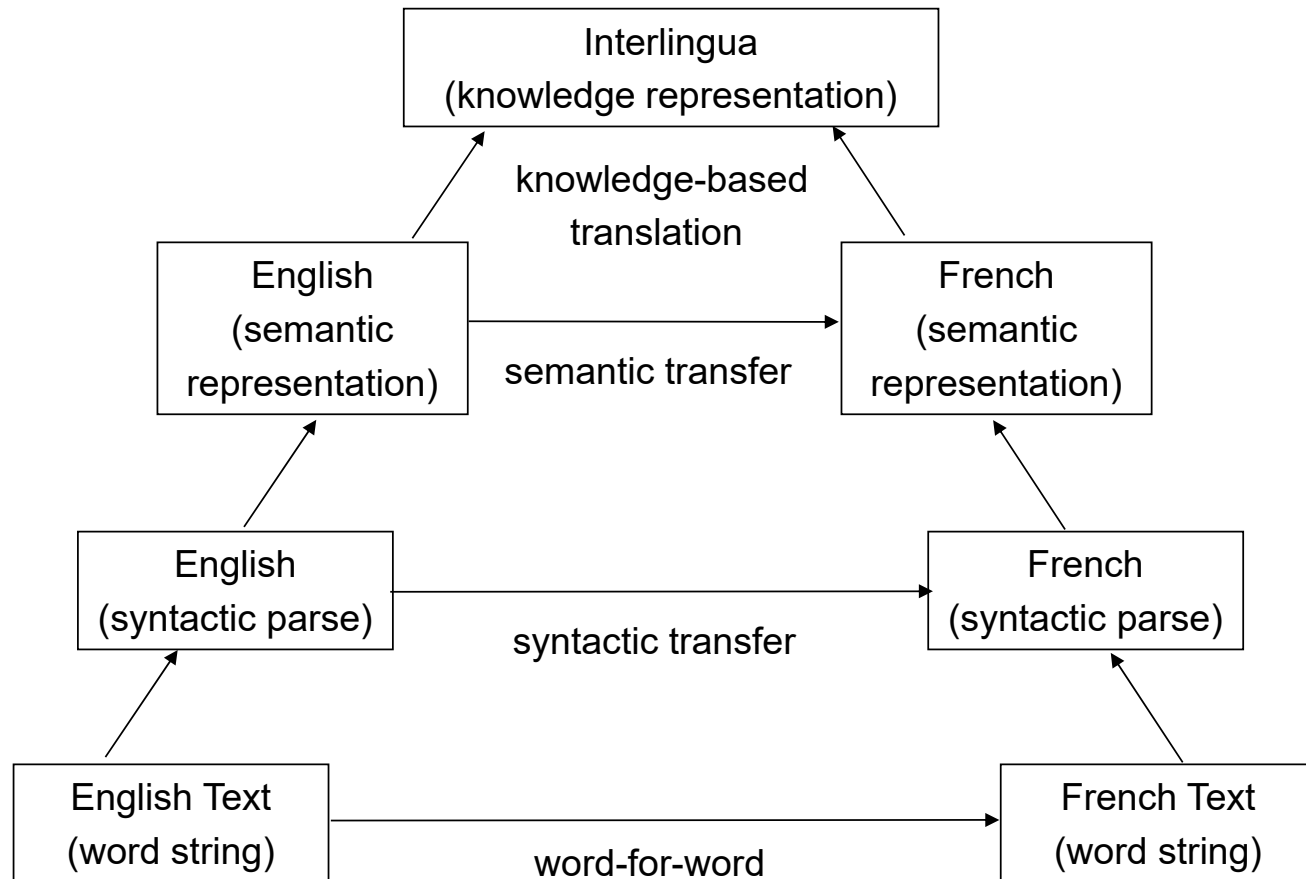
The Vauquois (1968) Triangle

Neural multilingual MT ?

Neural MT

Conventional statistical MT
(Transfer Approaches)

Direct Translation



Neural Machine Translation (1/3)

- The current de facto (standard) architecture for MT is the encoder-decoder transformer or sequence-to-sequence (RNN, LSTM and others) model
- For example, a basic RNN (recurrent neural network) version of encoder-decoder approach to machine translation

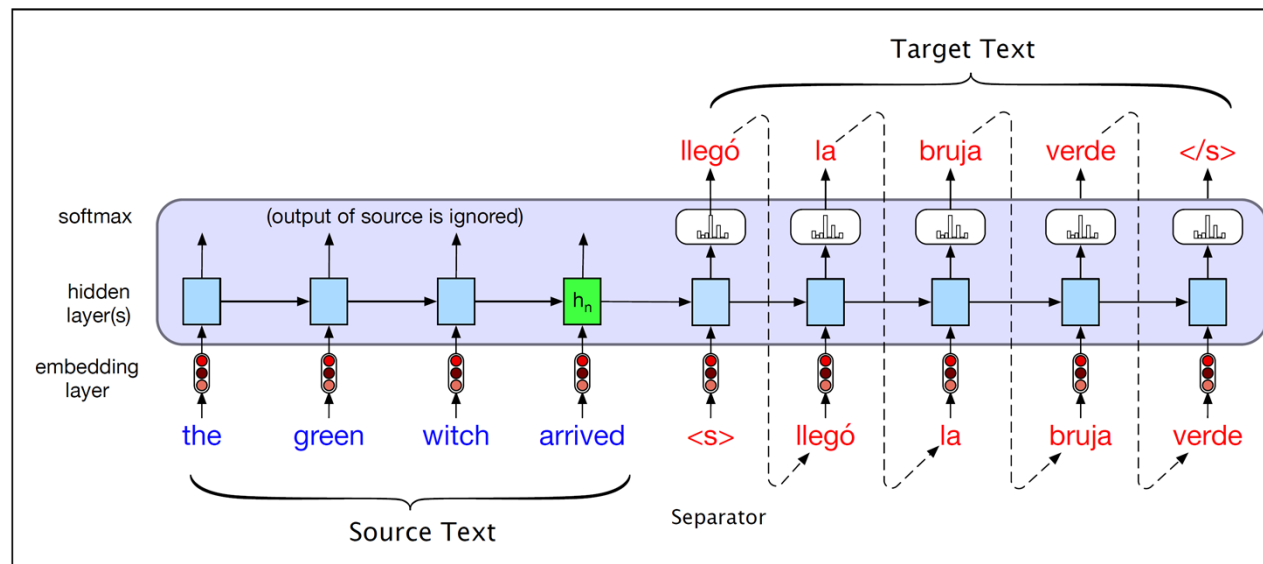


Figure 9.17 Translating a single sentence (inference time) in the basic RNN version of encoder-decoder approach to machine translation. Source and target sentences are concatenated with a separator token in between, and the decoder uses context information from the encoder's last hidden state.

Neural Machine Translation (2/3)

- A Transformer-based Encoder-Decoder MY Architecture

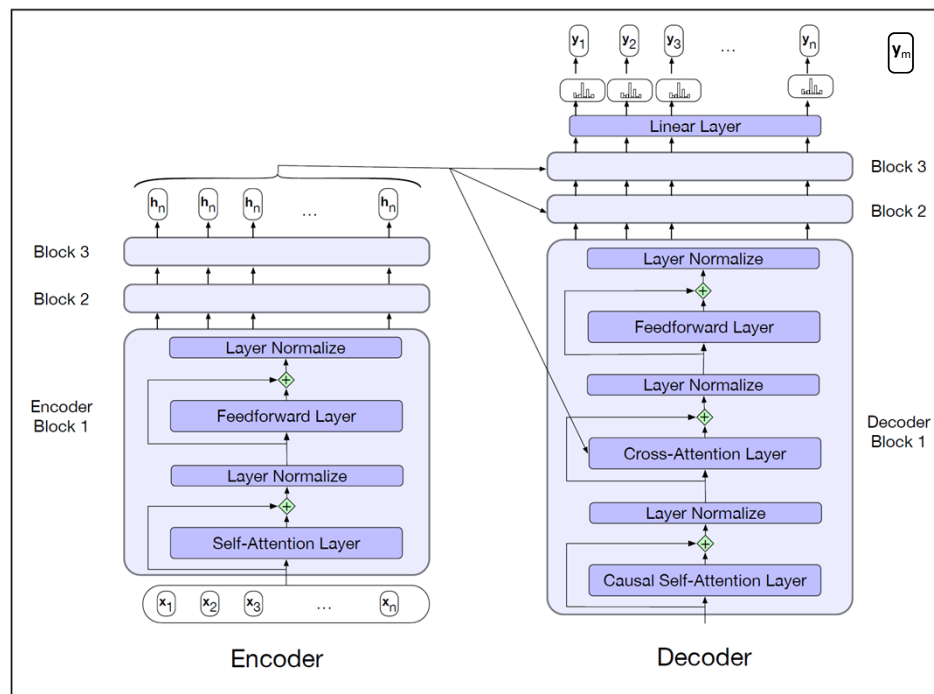


Figure 13.6 The transformer block for the encoder and the decoder. The final output of the encoder $\mathbf{H}^{enc} = \mathbf{h}_1, \dots, \mathbf{h}_T$ is the context used in the decoder. The decoder is a standard transformer except with one extra layer, the **cross-attention** layer, which takes that decoder output \mathbf{H}^{enc} and uses it to form its \mathbf{K} and \mathbf{V} inputs.

$$Q = W^Q \mathbf{H}^{dec[i-1]};$$

$$K = W^K \mathbf{H}^{enc};$$

$$V = W^V \mathbf{H}^{enc};$$

$$\text{CrossAttention}(Q, K, V)$$

$$= \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

The model generates the token sequence of the target language in an autoregressive and left-to-right manner

Neural Machine Translation (3/3)

- Training of NMT

- At training time the system is given a large set of parallel sentences (each sentence in a source language matched with a sentence in the target language), and learns to map source sentences into target sentences
- In practice, rather than using words (as in the example above), the sentences are into a sequence of subword tokens (tokens can be words, or subwords, or individual characters)

- The systems are then trained to maximize the probability of the sequence of tokens in the target language y_1, \dots, y_m given the sequence of tokens in the source language x_1, \dots, x_n :

$$P(y_1, \dots, y_m \mid x_1, \dots, x_n)$$

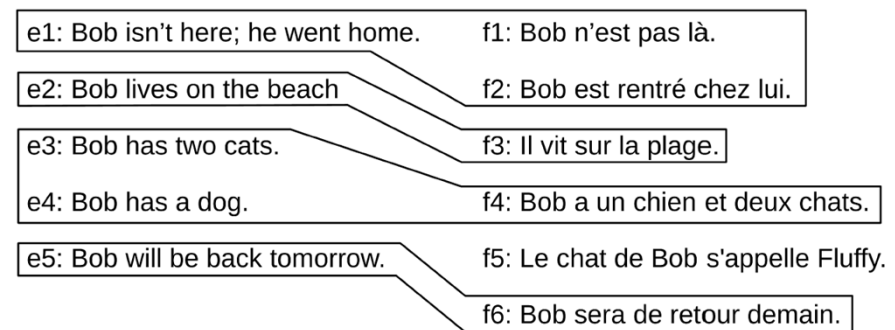


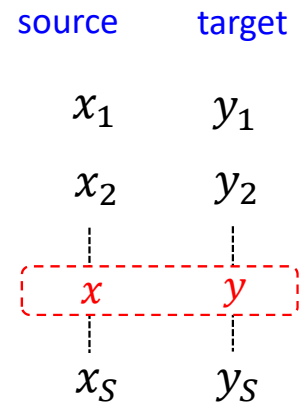
Figure 1: Sentence alignment takes sentences e_1, \dots, e_N and f_1, \dots, f_M and locates minimal groups of sentences which are translations of each other, in this case $(e_1)-(f_1, f_2)$, $(e_2)-(f_3)$, $(e_3, e_4)-(f_4)$, and $(e_5)-(f_6)$.

Creating Training Data for MT

- Machine translation models are trained on a parallel corpus, sometimes called a **bitext**, a text that appears in two (or more) languages
- Large numbers of parallel corpora are available, for example
 - **Europarl corpus**: extracted from the proceedings of the European Parliament, contains between 400,000 and 2 million sentences each from 21 European languages
 - **United Nations Parallel Corpus**: contains on the order of 10 million sentences in the six official languages of the United Nations (Arabic, Chinese, English, French, Russian, Spanish)
 - **OpenSubtitles corpus**: made from movie and TV subtitles
 - **ParaCrawl corpus**: 223 million sentence pairs between 23 EU languages and English extracted from the Common-Crawl dataset

Sentence Alignment

- Standard training corpora for MT come as aligned pairs of sentences
 - These sentence alignments must be created manually or automatically
- Typical Procedure for producing automatic sentence alignments
 - Step 1: a cost function that takes a span of source sentences and a span of target sentences and returns a score measuring how likely these spans are to be translations



cosine similarity between the embeddings of two spans of sentences

$$c(x, y) = \frac{(1 - \cos(\vec{x}, \vec{y}))nSents(x)nSents(y)}{\sum_{s=1}^S [1 - \cos(\vec{x}, \vec{y}_s)] + \sum_{s=1}^S [1 - \cos(\vec{x}_s, \vec{y})]}$$

number of sentences in y

cost function for aligning two spans of sentences

The denominator helps to normalize the similarities

- Step 2: an alignment algorithm that uses the cost function above to find a good alignment between the documents (with dynamic programming)

MT in Low-Resourced Situations

- The vast majority of the world's languages do not have large parallel training texts available
- An important ongoing research question is how to get good translation with lesser resourced languages
 - The resource problem can even be true for high resource languages when we need to translate into low-resourced domains
- Two typical methods to alleviate this problem
 - Data Augmentation with Backtranslation (回譯)
 - Multilingual MT

Data Augmentation with Backtranslation

- **Data augmentation** is a statistical technique for dealing with insufficient training data, by adding **new synthetic data** that is generated from the current natural data
- The most common data augmentation technique for machine translation is called backtranslation
- **Backtranslation** assumes that we have a larger amount of monolingual corpora in the target language
 - **Step 1:** Given a small parallel text (a bitext) in the source/target languages, We first use the bitext to train a MT system in the reverse direction: **a target-to-source MT system**
 - **Step 2:** Use the MT system trained in Step 1 to translate the monolingual target data to the source language
 - **Step 3:** Add this synthetic bitext (natural target sentences, aligned with MT-produced source sentences) to our training data, and retrain our source-to-target MT model

Multilingual MT

- Train **a single MT system** by giving it parallel sentences in many different pairs of languages (**one model fits all**)
 - That means we need to tell the system which language to translate from and to!
 - Namely, the system is told which language is the source language by adding a special token l_{source} to the encoder, and is added a special token l_{target} to the decoder to tell it what is the target language

$$H^{enc} = \text{encoder}(x, l_{source})$$

$$y_i = \text{decoder}(H^{enc}, l_{target}, y_1, \dots, y_{i-1})$$

- One advantage of a multilingual MT model
 - They can improve the translation of lower-resourced languages by drawing on information from a similar language in the training data that happens to have more resources

MT Evaluations

- Human Evaluations
 - The most accurate evaluations use human raters, such as online crowd-workers, to evaluate each translation along the several dimensions:
 - **Fluency**: Intelligibility, Clarity, Readability, Naturalness
 - **Adequacy (Fidelity)**: How much of the information in the source was preserved in the target
 - **Ranking**: Raters prefer which candidate translations?
 - **post-editing**: Taking the MT output and changing it minimally until raters feel good enough
- Automatic Evaluations
 - **Character F-score**: a good machine translation will tend to contain characters and words that occur in a human translation of the same sentence
 - **BLEU Score** (BiLingual Evaluation Understudy):
 - A function of the n -gram **word precision** over all the sentences combined with a brevity penalty computed over the corpus as a whole
 - Compute this n -gram precision for unigrams, bigrams, trigrams, and fourgrams and take the geometric mean

More on Automatic Evaluations

- More recent metrics use BERT or other embeddings to allow synonyms to match between the reference y and candidate \tilde{y}

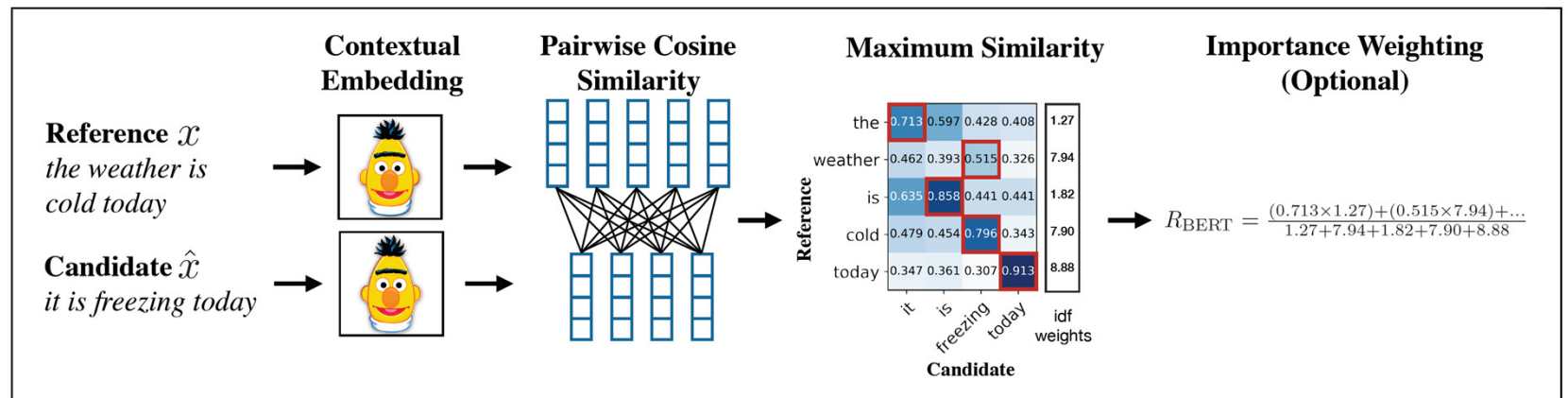


Figure 13.7 The computation of BERTSCORE recall from reference x and candidate \hat{x} , from Figure 1 in Zhang et al. (2020). This version shows an extended version of the metric in which tokens are also weighted by their idf values.

$$R_{\text{BERT}} = \frac{1}{\|\mathbf{x}\|_1} \sum_{x_i \in \mathbf{x}} \max_{\tilde{x}_j \in \tilde{\mathbf{x}}} \vec{x}_i \cdot \vec{\tilde{x}}_j$$

$$P_{\text{BERT}} = \frac{1}{\|\tilde{\mathbf{x}}\|_1} \sum_{\tilde{x}_j \in \tilde{\mathbf{x}}} \max_{x_i \in \mathbf{x}} \vec{x}_i \cdot \vec{\tilde{x}}_j$$

MT: Bias and Ethical Issues

- MT raises many ethical issues, for example
 - Consider MT systems translating from Hungarian (which has the gender neutral pronoun *ő*) or Spanish (which often drops pronouns) into English (in which pronouns are obligatory, and they have grammatical gender)

Hungarian (gender neutral) source	English MT output
ő egy ápoló	she is a nurse
ő egy tudós	he is a scientist
ő egy mérnök	he is an engineer
ő egy pék	he is a baker
ő egy tanár	she is a teacher
ő egy esküvőszervező	she is a wedding organizer
ő egy vezérigazgató	he is a CEO

Figure 13.8 When translating from gender-neutral languages like Hungarian into English, current MT systems interpret people from traditionally male-dominated occupations as male, and traditionally female-dominated occupations as female (Prates et al., 2019).

- One open problem is developing metrics for knowing what MT systems don't know for high-stakes tasks, e.g., **medical** and **legal** domains

Summary

- Machine translation is one of the most widely used applications of NLP, and the encoder-decoder model, **first developed for MT**, is a key tool that has applications throughout NLP
- Encoder-decoder networks are composed of **an encoder network** that takes an input sequence and creates a contextualized representation of it, the context. This context representation is then passed to **a decoder network** which generates a task-specific output sequence
- **Backtranslation** is a way of making use of monolingual corpora in the target language by running a pilot MT engine backwards to create synthetic bitexts

Appendix Material

Syntactic Transfer MT

- Parse the source text, then transfer the parse tree of the source text into a syntactic tree in the target language, and then generate the translation from this syntactic tree
 - Solve the problems of word ordering
- Problems
 - Syntactic ambiguity
 - The target syntax will likely mirror that of the source text

German: N V Adv
Ich esse gern (*I like to eat*)

English: I eat readily/gladly

Semantic Transfer MT

- Represent the meaning of the source sentence and then generate the translation from the meaning
 - Fix cases of syntactic mismatch
- Problems
 - Still be unnatural to the point of being unintelligible
 - Difficult to build the translation system for all pairs of languages

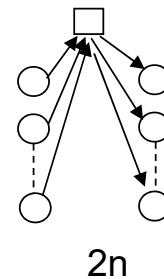
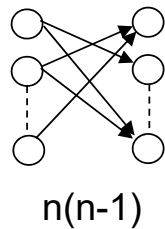
Spanish: La botella entró a la cueva flotando
(The bottle floated into the cave)

English: The bottle entered the cave floating

(In Spanish, the direction is expressed using the verb
and the manner is expressed with a separate phrase)

Knowledge-Based MT

- The translation is performed by way of a knowledge representation formalism called “interlingua”
 - Independence of the way particular languages express meaning
- Problems
 - Difficult to design an efficient and comprehensive knowledge representation formalism
 - Large amount of ambiguity needs to be solved to translate from a natural language to a knowledge representation language



Text Alignment: Definition

- Definition

- Align paragraphs, sentences or words in one language to paragraphs, sentences or words in another languages
 - Thus can learn which words tend to be translated by which other words in another language

bilingual dictionaries, MT , parallel grammars ...

- Is not part of MT process per se
 - But the obligatory first step for making use of multilingual text corpora

- Applications

- Bilingual lexicography
- Machine translation
- Multilingual information retrieval
- ...

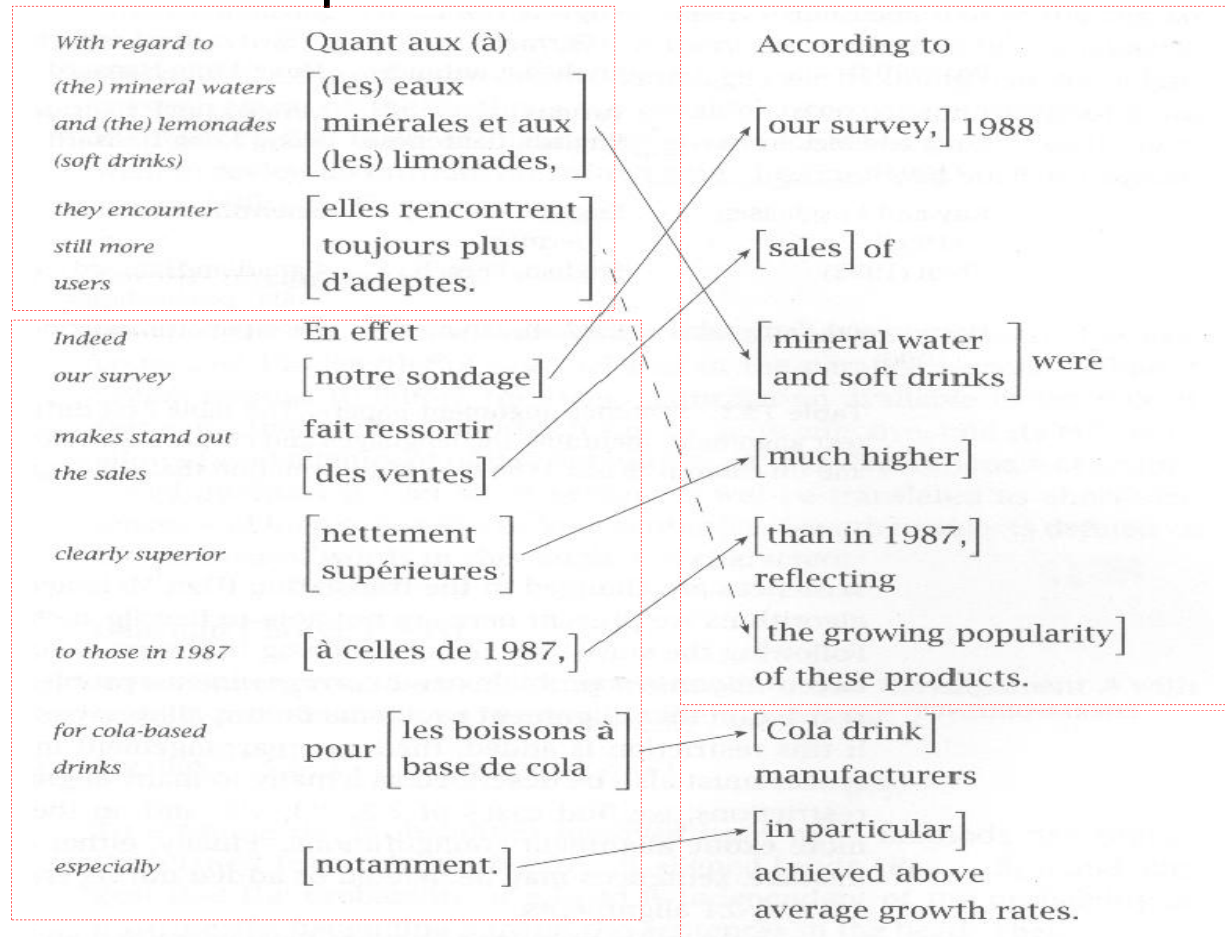
Text Alignment: Sources and Granularities

- Sources of Parallel texts or bitexts
 - Parliamentary proceedings (Hansards)
 - Newspapers and magazines
 - Religious and literary works
- Two levels of alignment
 - **Gross large scale alignment**
 - Learn which paragraphs or sentences correspond to which paragraphs or sentences in another language
 - **Word alignment**
 - Learn which words tend to be translated by which words in another language
 - The necessary step for acquiring a bilingual dictionary

} with less literal translation

Orders of word or sentence might not be preserved.

Text Alignment: Example 1



2:2 alignment

Figure 13.2 Alignment and correspondence. The middle and right columns show the French and English versions with arrows connecting parts that can be viewed as translations of each other. The italicized text in the left column is a fairly literal translation of the French text.

Text Alignment: Example 2

English	French	
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.	2:2 alignment
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.	1:1 alignment
Employment and investment levels also climbed.	L'emploi et les investissements ont également augmenté.	1:1 alignment
Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.	2:1 alignment

a bead/a sentence alignment

Studies show that around 90% of alignments are 1:1 sentence alignment.

Sentence Alignment (1/2)

- Crossing dependencies are not allowed here
 - Word ordering is preserved !
- Related work

Paper	Languages	Corpus	Basis
Brown et al. (1991c)	English, French	Canadian Hansard	# of words
Gale and Church (1993)	English, French, German	Union Bank of Switzerland reports	# of characters
Wu (1994)	English, Cantonese	Hong Kong Hansard	# of characters
Church (1993)	various	various (incl. Hansard)	4-gram signals
Fung and McKeown (1994)	English, Cantonese	Hong Kong Hansard	lexical signals
Kay and Röscheisen (1993)	English, French, German	Scientific American	lexical (not probabilistic)
Chen (1993)	English, French	Canadian Hansard EEC proceedings	lexical
Haruno and Yamazaki (1996)	English, Japanese	newspaper, magazines	lexical (incl. dictionary)

Sentence Alignment (2/2)

- Length-based
- Lexical-guided
- Offset-based

Sentence Alignment: Length-based method (1/9)

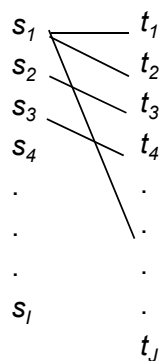
- **Rationale:** the short sentences will be translated as short sentences and long sentences as long sentences
 - Length is defined as the number of words or the number of characters

- **Approach 1** (Gale & Church 1993)

Union Bank of Switzerland (UBS) corpus
: English, French, and German

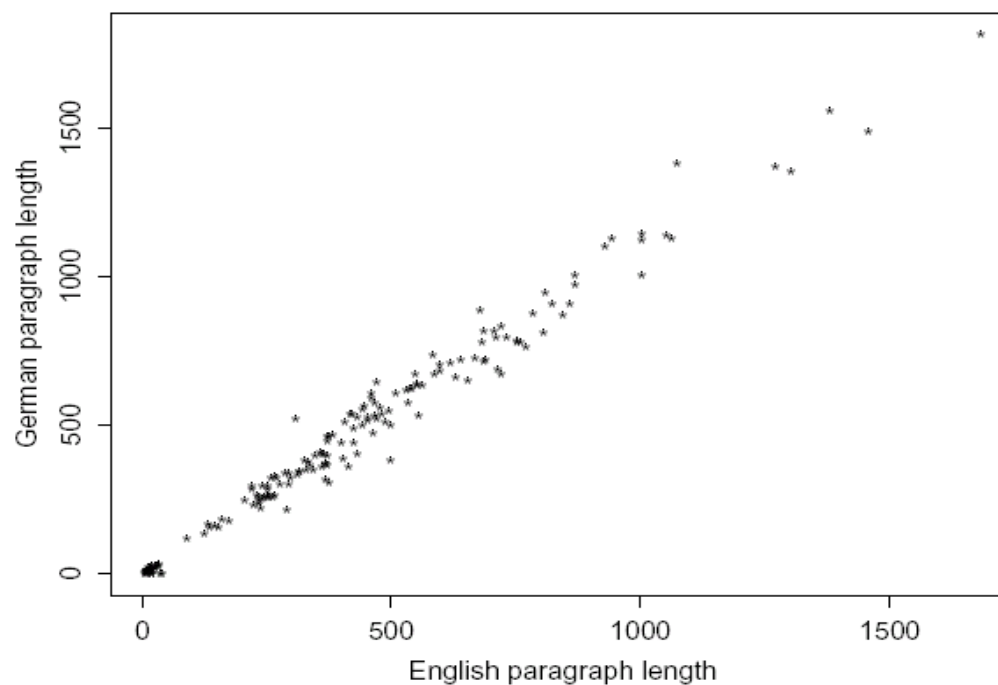
- **Assumptions**

- The paragraph structure was clearly marked in the corpus, confusions are checked by hand
- Lengths of sentences measured in characters
- **Crossing dependences** are not handled here
 - The order of sentences are not changed in the translation



Ignore the rich information available in the text.

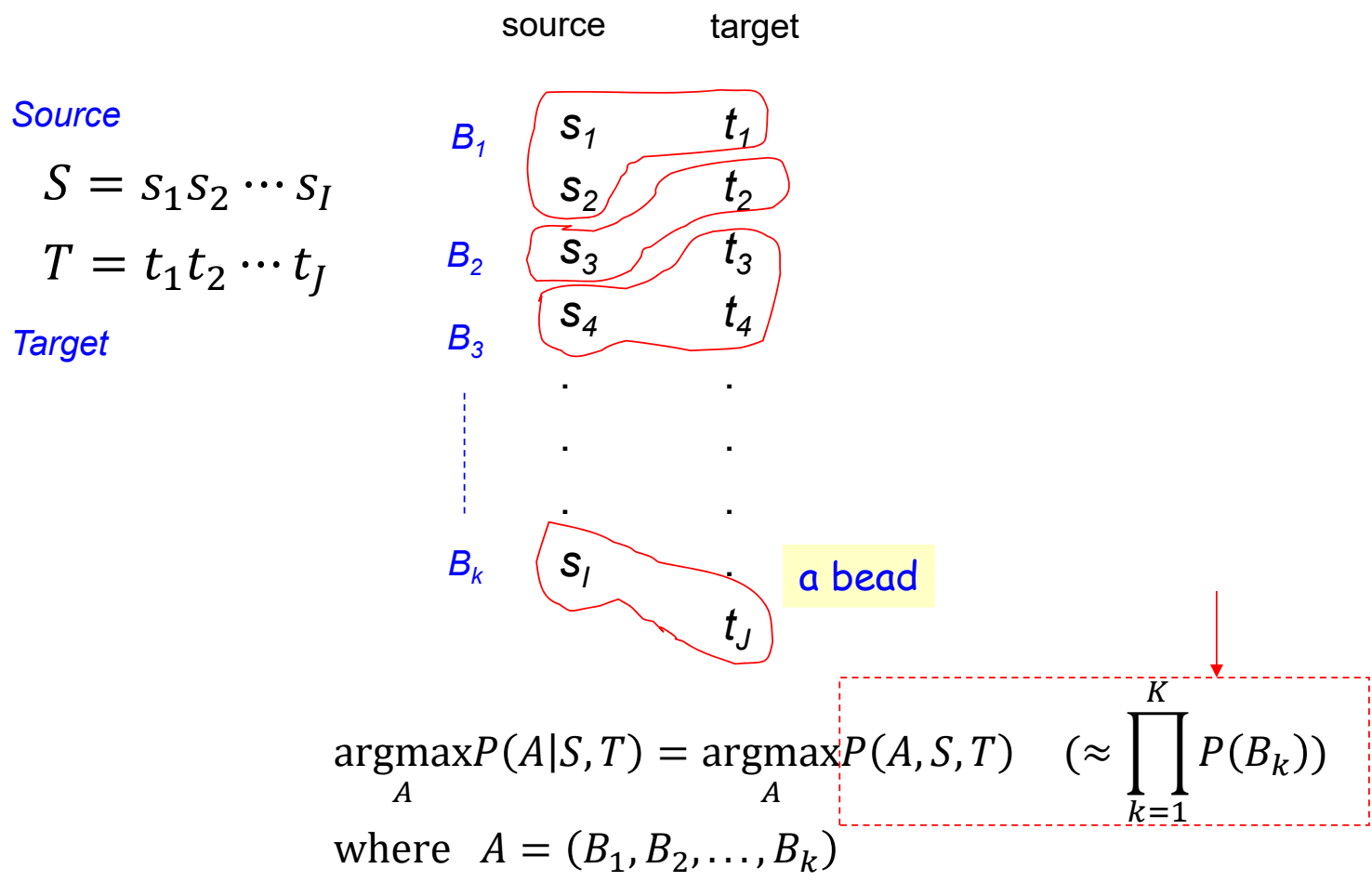
Sentence Alignment: Length-based method (2/9)



Most cases are
1:1 alignments.

Figure 1. The horizontal axis shows the length of English paragraphs, while the vertical scale shows the lengths of the corresponding German paragraphs. Note that the correlation is quite large (.991).

Sentence Alignment: Length-based method (3/9)



Sentence Alignment: Length-based method (4/9)


- Dynamic Programming

- The cost function (Distance Measure)

Bayes' Law

$$\begin{aligned} \text{cost}(\alpha \text{ align } l_1, l_2) &= -\log P(\alpha \text{ align} | \delta(l_1, l_2, \mu, s^2)) \\ &\approx -\log [P(\alpha \text{ align}) P(\delta(l_1, l_2, \mu, s^2) | \alpha \text{ align})] \end{aligned}$$

$-\log P(B_k)$ (with a red arrow pointing to the first term)



$$\delta(l_1, l_2, \mu, s^2) = (l_2 - l_1 \mu) / \sqrt{l_1 s^2}$$

Ratio of texts in two languages $\frac{L_2}{L_1} = \mu$ (with a red arrow pointing to μ)

$\delta(\cdot)$ is a distance measure which forms a normal distribution square difference of two paragraphs (with a red arrow pointing to the denominator)

- Sentence is the unit of alignment
- Statistically modeling of character lengths

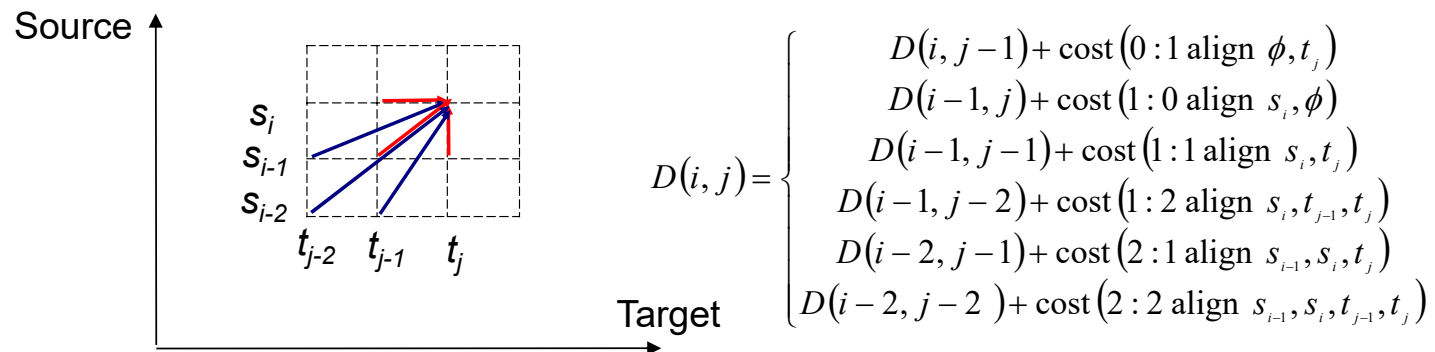
$$P(\delta(l_1, l_2, \mu, s^2) | \alpha \text{ align}) = 2(1 - \text{prob}(|\delta|))$$

The prob. distribution of standard normal distribution (with a red arrow pointing to the $\text{prob}(|\delta|)$ term)

Sentence Alignment: Length-based method (5/9)

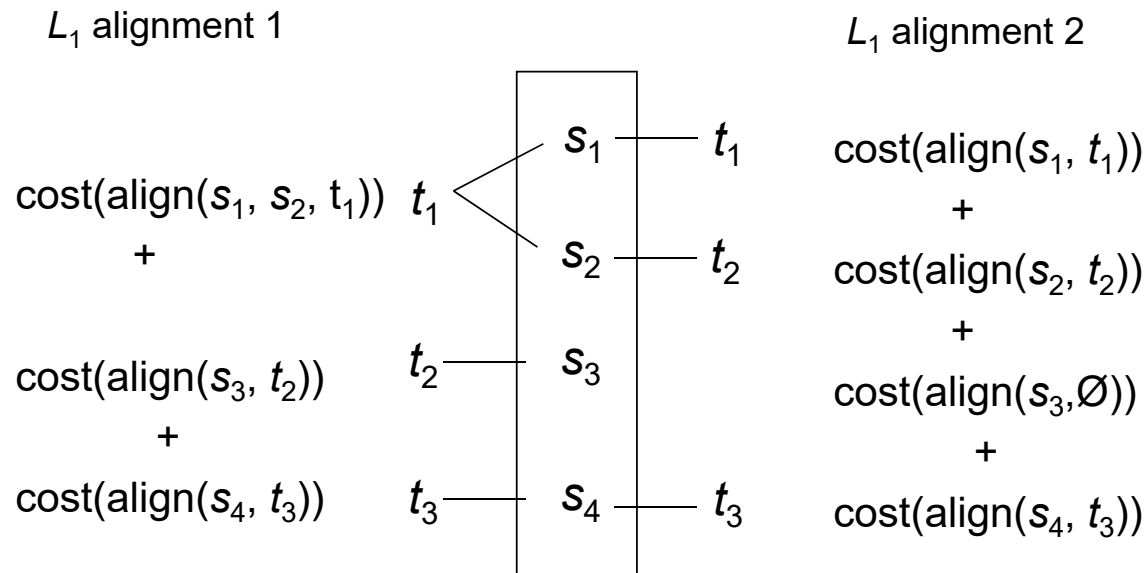
- The priori probability

Category	Frequency	Prob(match) Or $P(\alpha \text{ align})$
1-1	1167	0.89
1-0 or 0-1	13	0.0099
2-1 or 1-2	117	0.089
2-2	15	0.011
	1312	1.00



Sentence Alignment: Length-based method (6/9)

- A simple example



Sentence Alignment: Length-based method (7/9)

- The experimental results

category	English-French			English-German			total		
	N	err	%	N	err	%	N	err	%
1-0	8	8	100	5	5	100	13	13	100
1-1	542	14	2.6	625	9	1.4	1167	23	2.0
2-1	59	8	14	58	2	3.4	117	10	9
2-2	9	3	33	6	2	33	15	5	33
3-1	1	1	100	1	1	100	2	2	100
3-2	1	1	100	0	0	-	1	1	100

Sentence Alignment: Length-based method (8/9)

- 4% error rate was achieved
- **Problems:**
 - Can not handle noisy and imperfect input
 - E.g., OCR output or file containing unknown markup conventions
 - Finding paragraph or sentence boundaries is difficult
 - **Solution:** just align text (position) offsets in two parallel texts (Church 1993)
 - Questionable for languages with few cognates or different writing systems
 - E.g., English \longleftrightarrow Chinese

eastern European languages \longleftrightarrow Asian languages

Sentence Alignment: Length-based method (9/9)

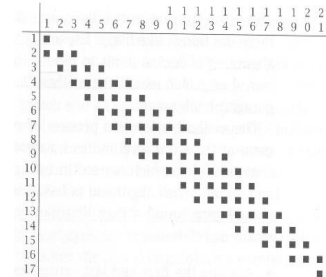
- **Approach 2** (Brown 1991)
 - Compare **sentence length in words** rather than characters
 - However, variance in number of words is greater than that of characters
 - EM training for the model parameters
- **Approach 3** (Wu 1994)
 - Apply the method of Gale and Church(1993) to a corpus of parallel English and Cantonese text
 - Also explore the use of lexical cues

Sentence Alignment: Lexical method (1/5)

- **Rationale:** the lexical information gives a lot of confirmation of alignments
 - Use a partial alignment of lexical items to induce the sentence alignment
 - That is, a partial alignment at the word level induces a maximum likelihood at the sentence level
 - The result of the sentence alignment can be in turn to refine the word level alignment

Sentence Alignment: Lexical method (2/5)

- Approach 1 (Kay and Röscheisen 1993)
 - First assume the first and last sentences of the text were align as the initial anchors
 - Form an envelope of possible alignments
 - Alignments excluded when sentences across anchors or their respective distance from an anchor differ greatly
 - Choose word pairs their distributions are similar in most of the sentences
 - Find pairs of source and target sentences which contain many possible lexical correspondences
 - The most reliable of pairs are used to induce a set of partial alignment (add to the list of anchors)



Iterations

Sentence Alignment: Lexical method (3/5)

- Approach 1
 - Experiments
 - On Scientific American articles
 - 96% coverage achieved after 4 iterations, the reminders is 1:0 and 0:1 matches
 - On 1000 Hansard sentences
 - Only 7 errors (5 of them are due to the error of sentence boundary detection) were found after 5 iterations
 - Problem
 - If a large text is accompanied with only endpoints for anchors, the pillow must be set to large enough, or the correct alignments will be lost
 - Pillow is treated as a constraint

Sentence Alignment: Lexical method (4/5)

- Approach 2 (Chen 1993)
 - Sentence alignment is done by constructing a simple word-to-word alignment
 - Best alignment is achieved by maximizing the likelihood of the corpus given the translation model
 - Like the method proposed by Gale and Church(1993), except that a translation model is used to estimate the cost of a certain alignment

$$\operatorname{argmax}_A P(A, S, T) \approx \prod_{k=1}^K P(B_k)$$

The translation model

$$\begin{aligned} -\log P(B_k) &= \text{cost}(\alpha \text{ align } l_1, l_2) \\ &\approx -\log[P(\alpha \text{ align})P(T(l_1, l_2)|\alpha \text{ align})] \end{aligned}$$


Sentence Alignment: Lexical method (5/5)

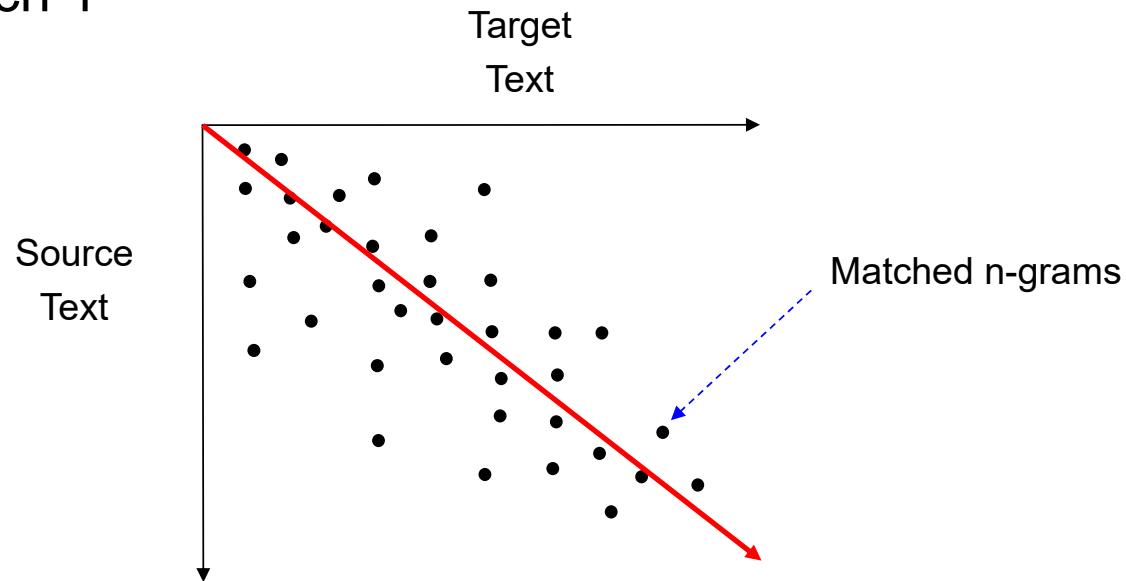
- Approach 3 (Haruno and Yamazaki, 1996)
 - Function words are left out and only content words are used for lexical matching
 - Part-of-speech taggers are needed
 - For short texts, [an on-line dictionary](#) is used instead of the finding of word correspondences adopted by Kay and Röscheisen (1993)

Offset Alignment (1/4)

- Perspective
 - Do not attempt to align beads of sentences but just align position offsets in two parallel texts
 - Avoid the influence of noises or confusions in texts
 - Can alleviate the problems caused by the absence of sentence markups
- Approach 1: (Church 1993)
 - Induce an alignment by cognates, proper nouns, numbers, etc.
 - **Cognate words**: words similar across languages
 - **Cognate words** share ample supply of identical character sequences between source and target languages
 - Use DP to find a alignment for the occurrence of matched character 4-grams along the diagonal line

Offset Alignment (2/4)

- Approach 1



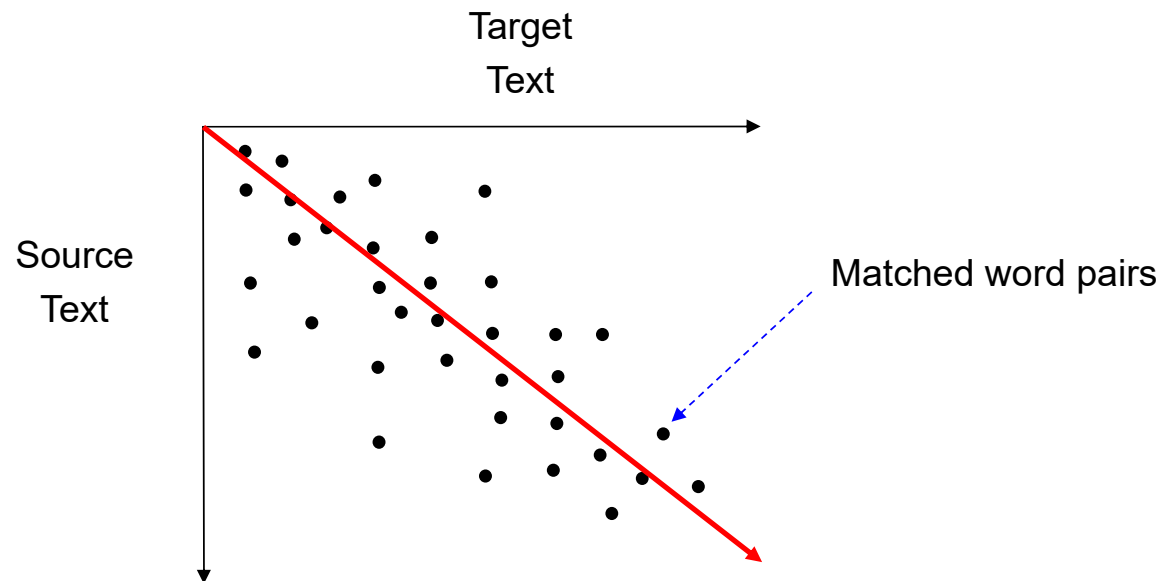
- Problem
 - Fail completely when language with different character sets (English \longleftrightarrow Chinese)

Offset Alignment (3/4)

- Approach 2: (Fung and McKeown 1993)
 - Two-stage processing
 - First stage (to infer a small bilingual dictionary)
 - For each word a signal is produced, as an arrival vector of integer number of words between each occurrence
 - E.g., word appears in offsets (1, 263, 267, 519) has an arrival vector (262,4,252)
 - Perform Dynamic Time Warping to match the arrival vectors of two English and Cantonese words to determine the similarity relations
 - Pairs of an English word and Cantonese word with very similar signals are retained in the dictionary
 - Properties
 - Genuinely language independent
 - Sensitive to lexical content

Offset Alignment (4/4)

- Approach 2: (Fung and McKeown 1993)
 - Second stage
 - Use DP to find a alignment for the occurrence of strongly-related word pairs along the diagonal line



Sentence/Offset Alignment: Summary

Paper	Languages	Corpus	Basis
Brown et al. (1991c)	English, French	Canadian Hansard	# of words
Gale and Church (1993)	English, French, German	Union Bank of Switzerland reports	# of characters
Wu (1994)	English, Cantonese	Hong Kong Hansard	# of characters
Church (1993)	various	various (incl. Hansard)	4-gram signals
Fung and McKeown (1994)	English, Cantonese	Hong Kong Hansard	lexical signals
Kay and Röscheisen (1993)	English, French, German	Scientific American	lexical (not probabilistic)
Chen (1993)	English, French	Canadian Hansard EEC proceedings	lexical
Haruno and Yamazaki (1996)	English, Japanese	newspaper, magazines	lexical (incl. dictionary)

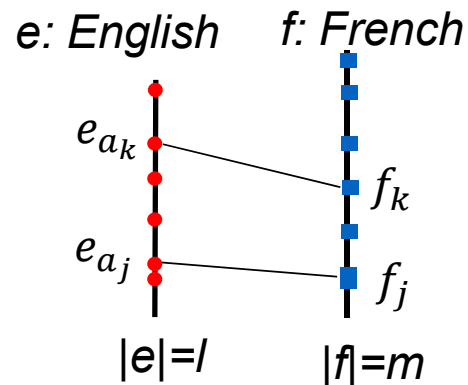
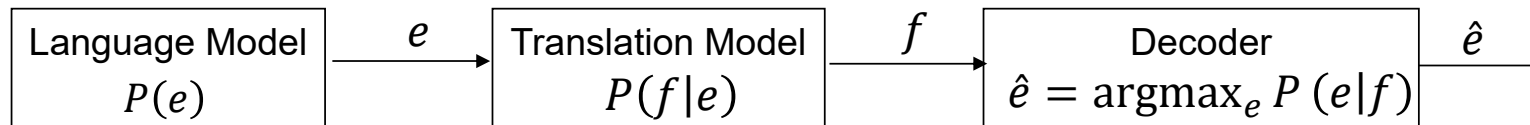
Table 13.1 Sentence alignment papers. The table lists different techniques for text alignment, including the languages and corpora that were used as a testbed and (in column “Basis”) the type of information that the alignment is based on.

Word Alignment

- The sentence/offset alignment can be extended to a word alignment
- Some criteria are then used to select aligned word pairs to include them into the bilingual dictionary
 - Frequency of word correspondences
 - Association measures
 -

Statistical Machine Translation (1/3)

- The noisy channel model



- Translation in sentence level
- **Assumptions:**
 - An English word can be aligned with multiple French words while each French word is aligned with at most English word
 - Independence of the individual word-to-word translations

Statistical Machine Translation (2/3)

- Three important components involved
 - Language model
 - Give the probability $p(e)$
 - Translation model

$$P(f|e) = \frac{1}{Z} \sum_{a_1=0}^l \cdots \sum_{a_m=0}^l \prod_{j=0}^m P(f_j | e_{a_j})$$

normalization constant all possible alignments translation probability
 (the English word that a French word f_j is aligned with)

- Decoder

$$\hat{e} = \operatorname{argmax}_e P(e|f) = \operatorname{argmax}_e \frac{p(e)p(f|e)}{p(f)} = p(e)p(f|e)$$

Statistical Machine Translation (3/3)

- EM Training

- E-step (Expectation)

$$Z_{w_f, w_e} = \sum_{(e, f) \text{ s.t. } w_e \in e, w_f \in f} P(w_f | w_e)$$

- M-step (Maximization)

↑
Number of times that w_e occurred in the English sentences while w_f in the corresponding French sentences

$$P(w_f | w_e) = \frac{Z_{w_f, w_e}}{\sum_v Z_{v, w_e}}$$

↑
 v : a given English word